Sample Complexity of Bayesian Optimal Dictionary Learning

Ayaka Sakata and Yoshiyuki Kabashima
Dep. of Computational Intelligence & Systems Science
Tokyo Institute of Technology
Yokohama 226-8502, Japan
Email: ayaka@sp.dis.titech.ac.jp, kaba@dis.titech.ac.jp

***Abstract*—We consider a learning problem of identifying a dictionary matrix $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ from a sample set of M dimensional vectors $\boldsymbol{Y} \in \mathbb{R}^{M\times P} = N^{-1/2}\boldsymbol{D}\boldsymbol{X} \in \mathbb{R}^{M\times P}$, where $\boldsymbol{X} \in \mathbb{R}^{N\times P}$ is a sparse matrix in which the density of non-zero entries is $0 < \rho < 1$. In particular, we focus on the minimum sample size $P_{\rm c}$ (sample complexity) necessary for perfectly identifying $\boldsymbol{D}$ of the optimal learning scheme when $\boldsymbol{D}$ and $\boldsymbol{X}$ are independently generated from certain distributions. By using the replica method of statistical mechanics, we show that $P_{\rm c} \sim O(N)$ holds as long as $\alpha = M/N > \rho$ is satisfied in the limit of $N \to \infty$. Our analysis also implies that the posterior distribution given $\boldsymbol{Y}$ is condensed only at the correct dictionary $\boldsymbol{D}$ when the compression rate α is greater than a certain critical value $\alpha_{\rm M}(\rho)$. This suggests that belief propagation may allow us to learn $\boldsymbol{D}$ with a low computational complexity using $O(N)$ samples.**

I. Introduction

The concept of sparse representations has recently attracted considerable attention from various fields in which the number of measurements is limited. Many real-world signals such as natural images are represented *sparsely* in Fourier/wavelet domains; in other words, many components vanish or are negligibly small in amplitude when the signals are represented by Fourier/wavelet bases. This empirical property is exploited in the signal recovery paradigm of compressed sensing (CS), thereby enabling the recovery of sparse signals from much fewer measurements than those estimated by the Nyquist-Shannon sampling theorem [1], [2], [3], [4].

In signal processing techniques for exploiting sparsity, signals are generally assumed to be described as linear combinations of a few dictionary atoms. Therefore, the effectiveness of this approach is highly dependent on the choice of dictionary, by which the objective signals appear sparse. A method for choosing an appropriate dictionary for sparse representation is *dictionary learning (DL)*, whereby the dictionary is constructed through a learning process from an available set of P training samples [5], [6], [7], [8].

The ambiguity of the dictionary is fatal in signal/data analysis after learning. Therefore, an important issue is the estimation of the *sample complexity*, i.e., the sample size P_c necessary for correct identification of the dictionary. In a seminal work, Aharon et al. showed that when the training set $\boldsymbol{Y} \in \mathbb{R}^{M\times P}$ is generated by a dictionary $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ and a sparse matrix $\boldsymbol{X} \in \mathbb{R}^{N\times P}$ (planted solution) as $\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{X}$, one can perfectly learn these if $P > P_{\rm c} = (k+1)_N C_k$ and k is sufficiently small, where k is the number of non-zero elements in each column of $\boldsymbol{X}$ [9]. Unfortunately, this bound becomes exponentially large in N for $k \sim O(N)$, which motivates us to improve the estimation. A recent rigorous study has shown that the mean squared error of recovered signals (per element) ϵ after learning can scale as $\epsilon \sim O\left(N\ln(kP)/P\right)$, which can be read as $P_{\rm c} \sim O(N\ln N)$ [10]. However, the expression still leads to a natural question: is the logarithmic factor intrinsic or not?

To answer this question, in this study, we evaluate the sample complexity of the optimal learning scheme defined for a given probabilistic model of dictionary learning. In a previous study, the authors assessed the sample complexity for a naive learning scheme: $\min_{\boldsymbol{D},\boldsymbol{X}} ||\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}||^2$ subj. to $||\boldsymbol{X}||_0 \leq NP\rho$ $(0 < \rho < 1)$, where $||\boldsymbol{X}||_0$ is the number of non-zero elements in $\boldsymbol{X}$ and $\boldsymbol{D}$ is enforced to be normalized appropriately. They used the replica method of statistical mechanics and found that $P_{\rm c} \sim O(N)$ holds when $\alpha = M/N$ is greater than a certain critical value $\alpha_{\rm naive}(\rho) > \rho$ [11]. However, the smallest possible $P_{\rm c}$ that can be obtained for $\alpha < \alpha_{\rm naive}(\rho)$ has not been clarified thus far. In this study, we show that $P_{\rm c} \sim O(N)$ holds in the entire region of $\alpha > \rho$ for the optimal learning scheme.

II. Problem setup

Let us suppose the following scenario of dictionary learning. Planted solutions, an $M \times N$ dictionary matrix $\boldsymbol{D} \in \mathbb{R}^{M\times N}$ and an $N \times P$ sparse matrix $\boldsymbol{X} \in \mathbb{R}^{N\times P}$, are independently generated from prior distributions

$$P(\boldsymbol{D}) = \frac{1}{\mathcal{N}_D}\prod_{i=1}^{N}\delta(\sum_{\mu} D_{\mu i}^2 - M), \tag{1}$$

$$P_\rho(\boldsymbol{X}) = \prod_{i,l} P_\rho(X_{il}) = \prod_{i,l}\Big\{(1-\rho)\delta(X_{il}) + \rho f(X_{il})\Big\}, \tag{2}$$

respectively, where $\mathcal{N}_D$ is a normalization constant and $\rho \in [0,1]$ is the rate of non-zero elements in $\boldsymbol{X}$. The distribution function $f(X)$ does not have a finite mass probability at the origin. The set of training samples $\boldsymbol{Y} \in \mathbb{R}^{M\times P}$, whose column vector corresponds to a training sample, is assumed to be given by the planted solutions as

$$\boldsymbol{Y} = \frac{1}{\sqrt{N}}\boldsymbol{D}\boldsymbol{X}, \tag{3}$$

where $1/\sqrt{N}$ is introduced for convenience in taking the large-system limit. A learner is required to infer $\boldsymbol{D}$ and $\boldsymbol{X}$ from $\boldsymbol{Y}$. Our aim is to evaluate the *minimum* value of the sample size P required for perfectly identifying $\boldsymbol{D}$ and $\boldsymbol{X}$.

III. Bayesian optimal learning

For mathematical formulation of our problem, let us denote the estimates of $\boldsymbol{D}$ and $\boldsymbol{X}$ yielded by an *arbitrary* learning scheme as $\hat{\boldsymbol{D}}(\boldsymbol{Y})$ and $\hat{\boldsymbol{X}}(\boldsymbol{Y})$. We evaluate the efficiency of the scheme using the mean squared errors (per element),

$$\mathrm{MSE}_D(\hat{\boldsymbol{D}}(\cdot))=\frac{1}{NM}\sum_{\boldsymbol{Y},\boldsymbol{D},\boldsymbol{X}}P_\rho(\boldsymbol{D},\boldsymbol{X},\boldsymbol{Y})||\boldsymbol{D}-\hat{\boldsymbol{D}}(\boldsymbol{Y})||^2 \quad (4)$$

$$\mathrm{MSE}_X(\hat{\boldsymbol{X}}(\cdot))=\frac{1}{NP}\sum_{\boldsymbol{Y},\boldsymbol{D},\boldsymbol{X}}P_\rho(\boldsymbol{D},\boldsymbol{X},\boldsymbol{Y})||\boldsymbol{X}-\hat{\boldsymbol{X}}(\boldsymbol{Y})||^2 \quad (5)$$

where $\boldsymbol{A}\cdot\boldsymbol{B}=\sum_{i,j}A_{ij}B_{ij}$ represents the inner product between two matrices of the same dimension $\boldsymbol{A}$ and $\boldsymbol{B}$, and $||\boldsymbol{A}||=\sqrt{\boldsymbol{A}\cdot\boldsymbol{A}}$ indicates the Frobenius norm of $\boldsymbol{A}$. We impose the normalization constraint $\sum_{\mu=1}^{M}(\hat{\boldsymbol{D}}(\boldsymbol{Y}))_{\mu i}^2=M$ for each column index $i=1,2,\ldots,N$ in order to avoid the ambiguity of the product $\hat{\boldsymbol{D}}(\boldsymbol{Y})\hat{\boldsymbol{X}}(\hat{Y})=\hat{\boldsymbol{D}}(\boldsymbol{Y})\boldsymbol{A}^{-1}\boldsymbol{A}\hat{\boldsymbol{X}}(\boldsymbol{Y})$ for an arbitrary invertible diagonal matrix $\boldsymbol{A}$ [1]. The joint distribution of $\boldsymbol{D}$, $\boldsymbol{X}$, and $\boldsymbol{Y}$ is given by

$$P_\rho(\boldsymbol{D},\boldsymbol{X},\boldsymbol{Y})=\delta(\boldsymbol{Y}-\frac{1}{\sqrt{N}}\boldsymbol{D}\boldsymbol{X})P(\boldsymbol{D})P_\rho(\boldsymbol{X}). \quad (6)$$

The perfect identification of $\boldsymbol{D}$ and $\boldsymbol{X}$ can be characterized by $\mathrm{MSE}_D=\mathrm{MSE}_X=0$. The following theorem offers a useful basis for answering our question.

Theorem 1. *For an arbitrary learning scheme, (4) and (5) are bounded from below as*

$$\mathrm{MSE}_D(\hat{\boldsymbol{D}}(\cdot))\geq 2-2\sum_{\boldsymbol{Y}}P_\rho(\boldsymbol{Y})\left(\frac{1}{N}\sum_{i=1}^{N}\frac{||(\langle\boldsymbol{D}\rangle_\rho)_i||}{\sqrt{M}}\right) \quad (7)$$

$$\mathrm{MSE}_X(\hat{\boldsymbol{X}}(\cdot))\geq\sum_{\boldsymbol{Y}}P_\rho(\boldsymbol{Y})\left(\frac{\langle\boldsymbol{X}\cdot\boldsymbol{X}\rangle_\rho}{NP}-\frac{\langle\boldsymbol{X}\rangle_\rho\cdot\langle\boldsymbol{X}\rangle_\rho}{NP}\right), \quad (8)$$

where $P_\rho(\boldsymbol{Y})=\sum_{\boldsymbol{D},\boldsymbol{X}}P_\rho(\boldsymbol{D},\boldsymbol{X},\boldsymbol{Y})$, and $\langle\cdot\rangle_\rho$ denotes the average over $\boldsymbol{D}$ and $\boldsymbol{X}$ according to the posterior distribution of $\boldsymbol{D}$ and $\boldsymbol{X}$ under a given $\boldsymbol{Y}$, $P_\rho(\boldsymbol{D},\boldsymbol{X}|\boldsymbol{Y})=P_\rho(\boldsymbol{D},\boldsymbol{X},\boldsymbol{Y})/P_\rho(\boldsymbol{Y})$. The equalities hold when the estimates satisfy

$$(\hat{\boldsymbol{D}}^{\mathrm{opt}}(\boldsymbol{Y}))_i=\sqrt{M}\frac{(\langle\boldsymbol{D}\rangle_\rho)_i}{||(\langle\boldsymbol{D}\rangle_\rho)_i||},\quad \hat{\boldsymbol{X}}^{\mathrm{opt}}(\boldsymbol{Y})=\langle\boldsymbol{X}\rangle_\rho, \quad (9)$$

[1]Additionally, $\hat{\boldsymbol{D}}(\boldsymbol{Y})\hat{\boldsymbol{X}}(\boldsymbol{Y})$ is invariant under any simultaneous permutations of columns in $\hat{\boldsymbol{D}}(\boldsymbol{Y})$ and rows in $\hat{\boldsymbol{X}}(\boldsymbol{Y})$, which yields an $N!$ degeneracy of an intrinsically identical solution. However, this does not influence the results of the current analysis since the number of degeneracy $N!$ is negligible in the saddle point assessment of $[P_\theta^n(\boldsymbol{Y})]_Y$ which scales exponentially in N^2.

where $(\boldsymbol{A})_i$ denotes the i-th column vector of matrix $\boldsymbol{A}$. We refer to (9) as the Bayesian optimal learning scheme [12].

Proof: By applying the Cauchy-Shwartz inequality and the minimization of the quadratic function to MSE_D and MSE_X, respectively, one can obtain (7)–(9) after inserting the expression

$$\sum_{\boldsymbol{D},\boldsymbol{X}}xP_\rho(\boldsymbol{D},\boldsymbol{X},\boldsymbol{Y})=P_\rho(\boldsymbol{Y})\sum_{\boldsymbol{D},\boldsymbol{X}}xP_\rho(\boldsymbol{D},\boldsymbol{X}|\boldsymbol{Y})=P_\rho(\boldsymbol{Y})\langle x\rangle_\rho$$

for $x=\boldsymbol{D}$ and $\boldsymbol{X}$ into (4). □

This theorem guarantees that when the setup of dictionary learning is characterized by (1)–(3), the estimates of (9) offer the *best possible learning performance* in the sense that (4) and (5) are minimized. As the perfect identification of $\boldsymbol{D}$ and $\boldsymbol{X}$ is characterized by $\mathrm{MSE}_D=\mathrm{MSE}_X=0$, our purpose is fulfilled by analyzing the performance of the Bayesian optimal learning scheme of (9).

IV. Analysis

For simplicity of calculation, let us set $f(X_{il})$ as the Gaussian distribution with mean 0 and variance σ_X^2, and σ_X^2 is set to unity for all numerical calculations later on. For generality, we consider cases in which the sparsity assumed by the learner, denoted as θ, can differ from the actual value ρ. When $\theta\neq\rho$, the estimates are given by $(\hat{\boldsymbol{D}}(\boldsymbol{Y}))_i=\sqrt{M}(\langle\boldsymbol{D}\rangle_\theta)_i/||(\langle\boldsymbol{D}\rangle_\theta)_i||$ and $\hat{\boldsymbol{X}}(\boldsymbol{Y})=\langle\boldsymbol{X}\rangle_\theta$ instead of (9). To evaluate MSE_D and MSE_X, we need to evaluate macroscopic quantities

$$q_D=\frac{1}{MN}[\langle\boldsymbol{D}\rangle_{\theta\cdot}\langle\boldsymbol{D}\rangle_\theta]_Y,\quad m_D=\frac{1}{MN}[\langle\boldsymbol{D}\rangle_{\theta\cdot}\langle\boldsymbol{D}\rangle_\rho]_Y \quad (10)$$

$$Q_X=\frac{1}{NP}[\langle\boldsymbol{X}\cdot\boldsymbol{X}\rangle_\theta]_Y \quad (11)$$

$$q_X=\frac{1}{NP}[\langle\boldsymbol{X}\rangle_{\theta\cdot}\langle\boldsymbol{X}\rangle_\theta]_Y,\quad m_X=\frac{1}{NP}[\langle\boldsymbol{X}\rangle_{\theta\cdot}\langle\boldsymbol{X}\rangle_\rho]_Y, \quad (12)$$

where $[\cdot]_Y=\sum_{\boldsymbol{Y}}P_\rho(\boldsymbol{Y})(\cdot)$. Note that (10)–(12) yield $\mathrm{MSE}_D\simeq 2-2m_D/\sqrt{q_D}$[2] and $\mathrm{MSE}_X=\rho\sigma_X^2+q_X-2m_X$.

Unfortunately, evaluating these is intrinsically difficult because it generally requires averaging the quantity

$$\frac{\sum_{\boldsymbol{D}^1,\boldsymbol{X}^1,\boldsymbol{D}^2,\boldsymbol{X}^2}P_\theta(\boldsymbol{Y},\boldsymbol{D}^1,\boldsymbol{X}^1)P_\theta(\boldsymbol{Y},\boldsymbol{D}^2,\boldsymbol{X}^2)(\boldsymbol{D}^1\cdot\boldsymbol{D}^2)}{\sum_{\boldsymbol{D}^1,\boldsymbol{X}^1,\boldsymbol{D}^2,\boldsymbol{X}^2}P_\theta(\boldsymbol{Y},\boldsymbol{D}^1,\boldsymbol{X}^1)P_\theta(\boldsymbol{Y},\boldsymbol{D}^2,\boldsymbol{X}^2)}$$
$$(=\langle\boldsymbol{D}\rangle_{\theta\cdot}\langle\boldsymbol{D}\rangle_\theta), \quad (13)$$

which includes summations over exponentially many terms in the denominator, with respect to $\boldsymbol{Y}$. One promising approach for avoiding this difficulty involves multiplying $P_\theta^n(\boldsymbol{Y})=(\sum_{\boldsymbol{X},\boldsymbol{D}}P_\theta(\boldsymbol{Y},\boldsymbol{D},\boldsymbol{X}))^n$ $(n=2,3,\ldots\in\mathbb{N})$ inside the operation of $[\cdot]_Y$ for canceling the denominator of (13), which makes the evaluation of a modified average

$$q_D(n)=\frac{1}{MN}\frac{[P_\theta^n(\boldsymbol{Y})\langle\boldsymbol{D}\rangle_{\theta\cdot}\langle\boldsymbol{D}\rangle_\theta]_Y}{[P_\theta^n(\boldsymbol{Y})]_Y} \quad (14)$$

[2] Naive computation requires us to assess a column-wise overlap $C_{D,i}=M^{-1/2}[(\langle\boldsymbol{D}\rangle_\rho)_i(\langle\boldsymbol{D}\rangle_\theta)_i/\sqrt{(\langle\boldsymbol{D}\rangle_\theta)_i\cdot(\langle\boldsymbol{D}\rangle_\theta)_i}]_Y$ for each column index $i=1,2,\ldots,N$. However, the law of large numbers and the statistical uniformity allow the simplification of $C_{D,i}\to m_D/\sqrt{q_D}$ as $M=\alpha N$ tends to infinity.

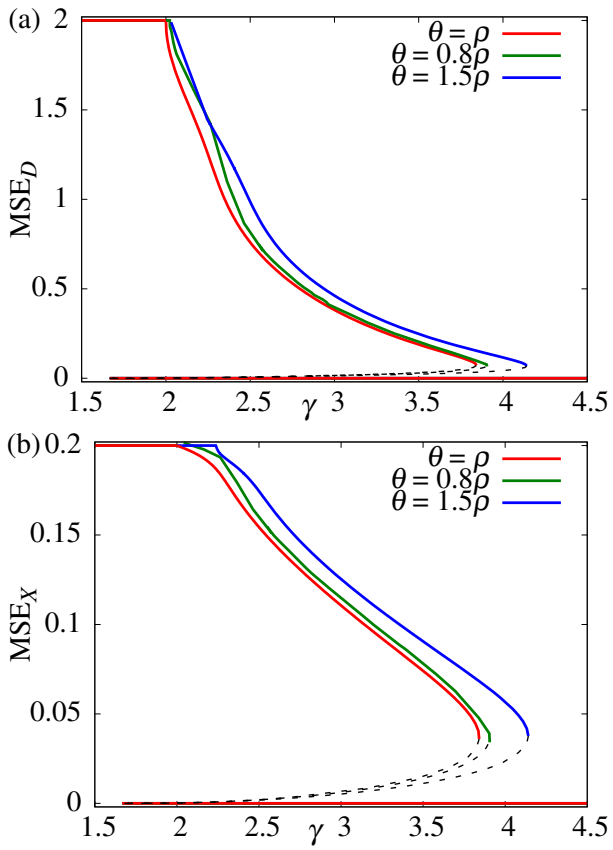

Fig. 1. γ-dependence of (a) MSE_D and (b) MSE_X at $\alpha = 0.5$, $\rho = 0.2$. Plots for $\theta = \rho$, 0.8ρ, and 1.5ρ show the optimality of the correct parameter choice $\theta = \rho$. Broken curves represent locally unstable branches, which are thermodynamically irrelevant.

feasible via the saddle point assessment of $[P_\theta^n(\boldsymbol{Y})]_Y$ for $N, M, P \to \infty$, keeping $\alpha = M/N$ and $\gamma = P/N$ as $O(1)$. Furthermore, the resulting expression is likely to hold for $n \in \mathbb{R}$ as well. Therefore, we evaluate q_D using the formula $q_D = \lim_{n\to 0} q_D(n)$ with the expression, and similarly, for m_D, Q_X, q_X, and m_X. This procedure is often termed the *replica method* [13], [14]. Under the replica symmetric ansatz, which assumes that the dominant saddle point in the evaluation is invariant under any permutation of replica indices $a = 1, 2, \ldots, n$, the assessment is reduced to evaluating the extremum of the free entropy (density) function

$$\phi = \gamma\Big(\frac{\hat{Q}_X Q_X + \hat{q}_X q_X}{2} - \hat{m}_X m_X + \langle\langle \ln \Xi_X \rangle\rangle\Big) + \frac{\alpha}{2}\Big(\hat{Q}_D + \hat{q}_D q_D - 2\hat{m}_D m_D - \ln(\hat{Q}_D + \hat{q}_D) + \frac{\hat{q}_D + \hat{m}_D^2}{\hat{Q}_D + \hat{q}_D}\Big) - \frac{\alpha\gamma}{2}\Big\{\frac{q_D q_X - 2m_D m_X + \rho\sigma_X^2}{Q_X - q_D q_X} + \ln(Q_X - q_D q_X)\Big\}, \quad (15)$$

where $\hat{\sigma}_X = 1 + (\hat{Q}_X + \hat{q}_X)\sigma_X^2$,

$$\Xi_X = (1-\theta) + \frac{\theta}{\sqrt{\hat{\sigma}_X}} \exp\Big(\frac{\sigma_X^2(\sqrt{\hat{q}_X} z + \hat{m}_X X^0)^2}{2\hat{\sigma}_X}\Big) \equiv (1-\theta) + \Xi_X^+, \quad (16)$$

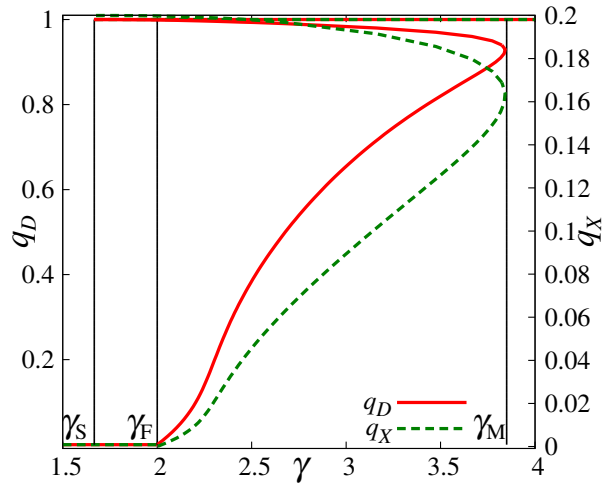

Fig. 2. γ-dependence of q_D (left axis) and q_X (right axis) for $\alpha = 0.5$ and $\rho = 0.2$.

and $\langle\langle\cdot\rangle\rangle$ denotes the average over X and z, which are distributed according to $P_\rho(X)$ and a Gaussian distribution with mean zero and variance 1, respectively. The extremized value[3] of ϕ, ϕ^*, is related to the average log-likelihood (density) of $\boldsymbol{Y}$ as $N^{-2}\sum_{\boldsymbol{Y}} P_\rho(\boldsymbol{Y}) \ln P_\theta(\boldsymbol{Y}) = \lim_{n\to 0}(\partial/\partial n)\left\{N^{-2}\ln[P_\theta^n(\boldsymbol{Y})]_Y\right\} = \phi^* + \text{constant}$.

In Fig. 1, (a) MSE_D and (b) MSE_X for $\theta = \rho = 0.2$ are plotted versus γ together with those for $\theta = 0.8\rho$ and $\theta = 1.5\rho$. At $\theta = \rho$, MSE_D and MSE_X of thermodynamically relevant branches have minimum values in the entire γ region, while a branch of solution characterized by $\mathrm{MSE}_D = \mathrm{MSE}_X = 0$ is shared by the three parameter sets. This supports the optimality of the correct parameter choice of $\theta = \rho$, and therefore, we hereafter focus our analysis on this case to estimate the minimum value of γ for the perfect learning, $\mathrm{MSE}_D = \mathrm{MSE}_X = 0$. At $\theta = \rho$, the relationships $m_D = q_D$, $m_X = q_X$, and $Q_X = \rho$ hold from (10)–(12), and the extremum problem is reduced to

$$q_D = \frac{\hat{q}_D}{1+\hat{q}_D}, \quad q_X = \left\langle\left\langle\left(\frac{\Xi_X^+}{\Xi_X}\frac{\sqrt{\hat{q}_X} z + \hat{q}_X X^0}{\hat{\sigma}_X^2}\right)^2\right\rangle\right\rangle, \quad (17)$$

where $\hat{q}_D$ and $\hat{q}_X$ are given by

$$\hat{q}_X = \frac{\alpha q_D}{\rho\sigma_X^2 - q_D q_X}, \quad \hat{q}_D = \frac{\gamma q_X}{\rho\sigma_X^2 - q_D q_X}. \quad (18)$$

The other variables are provided as $\hat{Q}_D = 1$, $\hat{Q}_X = 0$, $\hat{m}_X = \hat{q}_X$, and $\hat{m}_D = \hat{q}_D$.

V. Results

A. Actual solutions

Fig. 2 plots q_D and q_X versus γ for $\alpha = 0.5$ and $\rho = 0.2$. As shown in the figure, the solutions of q_D and q_X given by (17) are classified into three types: $q_D = 1$, $q_X = \rho\sigma_X^2$,

[3]When multiple extrema exist, the maximum value among them should be chosen as long as no consistency condition is violated.

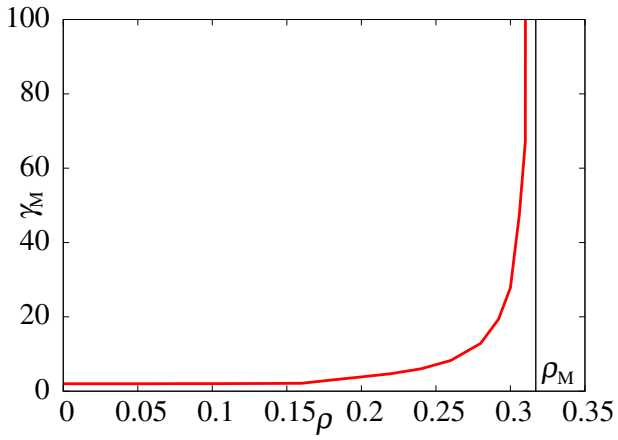

Fig. 3. $\gamma_{\rm M}$ versus ρ for $\alpha = 0.5$.

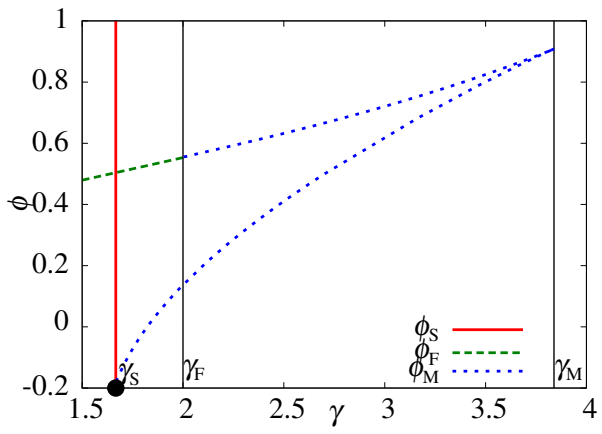

Fig. 4. γ-dependence of ϕ for $\alpha = 0.5$ and $\rho = 0.2$. $\phi_{\rm S}$ diverges positively for $\gamma > \gamma_{\rm S} = \alpha/(\alpha - \rho) = 1.666\ldots$.

$q_D = q_X = 0$, and $0 < q_D < 1$, $0 < q_X < \rho\sigma_X^2$. The first one yields $\mathrm{MSE}_D = \mathrm{MSE}_X = 0$, indicating the correct identification of $\boldsymbol{D}$ and $\boldsymbol{X}$, and hence, we name it the *success solution*. The second one is referred to as the *failure solution* because it yields $\mathrm{MSE}_D = 2$ and $\mathrm{MSE}_X = \rho\sigma_X^2$, which indicates complete failure of the learning of $\boldsymbol{D}$ and $\boldsymbol{X}$. The third one yields finite MSE_D and MSE_X, $0 < \mathrm{MSE}_D < 2$, $0 < \mathrm{MSE}_X < \rho\sigma_X^2$, and we term it the *middle solution*.

1) Success solution: When the expression

$$\delta\Big(\boldsymbol{Y} - \frac{\boldsymbol{D}\boldsymbol{X}}{\sqrt{N}}\Big) = \lim_{\tau\to+0}\Big(\frac{1}{\sqrt{2\pi\tau}}\Big)^{MP}\exp\Big(-\frac{||\boldsymbol{Y} - \frac{1}{\sqrt{N}}\boldsymbol{D}\boldsymbol{X}||^2}{2\tau}\Big), \tag{19}$$

is used, the success solution of q_D and q_X behaves as $(\rho\sigma_X^2 - q_X)/\tau = \chi_X$ and $(1 - q_D)/\tau = \chi_D$ while $\hat{q}_X$ and $\hat{q}_D$ scale as $\hat{q}_X = \hat{\theta}_X/\tau$ and $\hat{q}_D = \hat{\theta}_D/\tau$. By substituting them into the equations of q_D and q_X, they are given by

$$\chi_X = \frac{\rho\gamma}{g}, \quad \chi_D = \frac{\alpha}{\rho\sigma_X^2 g}, \quad \hat{\theta}_X = \frac{\rho}{\chi_X}, \quad \hat{\theta}_D = \frac{1}{\chi_D}, \tag{20}$$

where $g = (\alpha - \rho)\gamma - \alpha$. χ_X and χ_D must be positive by definition, and hence, the success solution exists for

$$\gamma > \frac{\alpha}{\alpha - \rho} \equiv \gamma_{\rm S} \tag{21}$$

only when $\alpha > \rho$.

2) Failure solution: The failure solution $q_D = q_X = 0$ appears at $0 \le \gamma < \gamma_{\rm F}$ as a locally stable solution. When q_D and q_X are sufficiently small, they are expressed as

$$q_X = \rho\sigma_X^2\alpha q_D + O(q^2), \quad q_D = \frac{\gamma q_X}{\rho\sigma_X^2} + O(q^2), \tag{22}$$

where $O(q^2)$ denotes the higher-order terms over second-order with respect to q_D and q_X. These expressions indicate that when

$$\gamma > \alpha^{-1} \equiv \gamma_{\rm F}, \tag{23}$$

the local stability of $q_D = q_X = 0$ is lost. As shown in Fig. 2, the failure solution vanishes at $\gamma_{\rm F} = 2.0$ for $\alpha = 0.5$.

3) Middle solution: We define $\gamma_{\rm M}$ over which the middle solution with $0 < q_D < 1$ and $0 < q_X < \rho\sigma_X^2$ disappears, denoted as a vertical line in Fig. 2, which is provided as $\gamma_{\rm M} = 3.841\ldots$ for the parameter choice of $(\alpha, \rho) = (0.5, 0.2)$. The value of $\gamma_{\rm M}$ depends on (α, ρ), as shown in Fig. 3. This figure indicates that $\gamma_{\rm M}$ diverges at $\rho_{\rm M} = 0.317\ldots$ for $\alpha = 0.5$. The relation between $\rho_{\rm M}$ and α, denoted as $\rho_{\rm M}(\alpha)$ (or $\alpha_{\rm M}(\rho)$), generally accords with the critical condition that belief propagation (BP)-based signal recovery using the correct prior starts to be involved with multiple fixed points for the signal reconstruction problem of compressed sensing [16] in which the correct dictionary $\boldsymbol{D}$ is provided in advance.

BP is also a potential algorithm for practically achieving the learning performance predicted by the current analysis because it is known that macroscopic behavior *theoretically* analyzed by the replica method can be confirmed *experimentally* for single instances by BP for many other systems [16], [17], [18]. The fact that only the success solution exists for $\gamma > \gamma_{\rm M}$ implies that one may be able to perfectly identify the correct dictionary $\boldsymbol{D}$ with a computational cost of *polynomial* order in N utilizing BP, without being trapped by other locally stable solutions, for $\alpha > \alpha_{\rm M}(\rho)$.

B. Free entropy density

There are three extrema of the free entropy (density), $\phi_{\rm S}$, $\phi_{\rm F}$, and $\phi_{\rm M}$, corresponding to the success solution, failure solution, and middle solution, respectively. Among them, the thermodynamically dominant solution that provides the correct evaluations of q_D and q_X is the one for which the value of free entropy is the largest. Fig. 4 plots $\phi_{\rm S}$, $\phi_{\rm F}$, and $\phi_{\rm M}$ versus γ for $\alpha = 0.5$, $\rho = 0.2$, where $\gamma_{\rm S} = 1.666\ldots$ and $\gamma_{\rm F} = 2.0$. In particular, functional forms of $\phi_{\rm S}$ and $\phi_{\rm F}$ are given by

$$\phi_{\rm S} = \lim_{\tau\to+0}\frac{1}{2}\Big[g\Big\{\ln\big(\frac{g}{\tau}\big) - 1\Big\} - \alpha\gamma\ln(\alpha\gamma) + \alpha\Big\{1 - \ln\Big(\frac{\rho\sigma_X^2}{\alpha}\Big)\Big\} + \gamma\rho(\ln\gamma - \ln\sigma_X^2)\Big] - \gamma H(\rho) \tag{24}$$

$$\phi_{\rm F} = \frac{1}{2}\{-\alpha\gamma(1 + \log\rho\sigma_X^2) + \alpha\}, \tag{25}$$

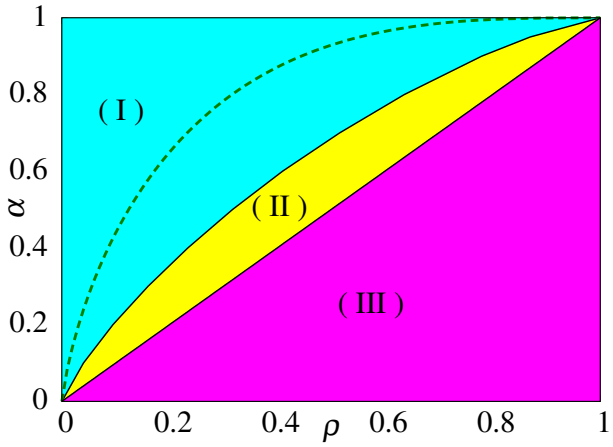

Fig. 5. Phase diagram on $\alpha - \rho$ plane. The dashed curve in the area of (I) is the result of [11].

where $\tau \to +0$ originates from the expression of (19) and $H(\rho) = -(1-\rho)\log(1-\rho) - \rho\log(\rho)$. Further, (24) shows that $\phi_{\rm S}$ diverges positively for $g = (\alpha - \rho)\gamma - \alpha > 0$, which guarantees that the success solution is always thermodynamically dominant for $\gamma > \gamma_{\rm S} = \alpha/(\alpha - \rho)$ as ϕ of other solutions is kept finite. This leads to the conclusion that the sample complexity of the Bayesian optimal learning is $P_{\rm c} = N\gamma_{\rm S}$, which is guaranteed as $O(N)$ as long as $\alpha > \rho$. This is the main consequence of the present study.

Fig. 5 plots the phase diagram in the $\alpha - \rho$ plane. The union of the regions (I) and (II) represents the condition that the sample complexity $P_{\rm c}$ is $O(N)$, while the full curve of the upper boundary of (II) denotes $\alpha_{\rm M}(\rho)$ above which BP is expected to work as an efficient learning algorithm. Dictionary learning is impossible in the region of (III). The critical condition $\alpha_{\rm naive}(\rho)$ above which the naive learning scheme of [11] can perfectly identify the planted solution by $O(N)$ samples is drawn as the dashed curve for comparison. The considerable difference between $\alpha_{\rm naive}(\rho)$ and ρ (or even $\alpha_{\rm M}(\rho)$) indicates the significance of using adequate knowledge of probabilistic models in dictionary learning.

VI. Summary

In summary, we assessed the minimum sample size required for perfectly identifying a planted solution in dictionary learning (DL). For this assessment, we derived the optimal learning scheme defined for a given probabilistic model of DL following the framework of Bayesian inference. Unfortunately, actually evaluating the performance of the Bayesian optimal learning scheme involves an intrinsic technical difficulty. For resolving this difficulty, we resorted to the replica method of statistical mechanics, and we showed that the sample complexity can be reduced to $O(N)$ as long as the compression rate α is greater than the density ρ of non-zero elements of the sparse matrix. This indicates that the performance of a naive learning scheme examined in a previous study [11] can be improved significantly by utilizing the knowledge of adequate probabilistic models in DL. It was also shown that when α is greater than a certain critical value $\alpha_{\rm M}(\rho)$, the macroscopic state corresponding to perfect identification of the planted solution becomes a unique candidate for the thermodynamically dominant state. This suggests that one may be able to learn the planted solution with a computational complexity of polynomial order in N utilizing belief propagation for $\alpha > \alpha_{\rm M}(\rho)$.

– Note added: After completing this study, the authors became aware that [19] presents results similar to those presented in this paper, where an algorithm for dictionary learning/calibration is independently developed on the basis of belief propagation.

Acknowledgment

This work was partially supported by a Grant-in-Aid for JSPS Fellow No. 23–4665 (AS) and KAKENHI Nos. 22300003 and 22300098 (YK).

References

[1] J.-L. Starck, F. Murtagh , and J. M. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity* (Cambridge Univ. Press, New York, 2010).
[2] H. Nyquist, *Certain topics in telegraph transmission theory*, Trans. AIEE **47** (2), pp. 617–644 (1928).
[3] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (4), pp. 1289–1306 (2006).
[4] E. J. Candès, and T Tao, *Decoding by Linear Programming*, IEEE Trans. Inform. Theory **51** (12), pp. 4203–4215 (2005).
[5] B. A. Olshausen and D. J. Field, *Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?*, Vision Res. **37** (23), pp. 3311–3325 (1997).
[6] R. Rubinstein, A. M. Bruckstein, and M. Elad, *Dictionaries for Sparse Representation Modeling*, Proc. of IEEE **98** (6), pp. 1045–1057 (2010).
[7] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, (Springer-Verlag, New York, 2010).
[8] S. Gleichman, and Y. C. Eldar, *Blind Compressed Sensing*, IEEE Inform. Theory **57**, pp. 6958–6975 (2011).
[9] M. Aharon, M. Elad, and A. M. Bruckstein, *On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them*, Linear Algebra and its Applications **416** (1), pp. 48–67 (2006).
[10] D. Vainsencher, S. Mannor, and A. M. Bruckstein, *The Sample Complexity of Dictionary Learning*, Journal of Machine Learning Research **12**, pp. 3259–3281 (2011).
[11] A. Sakata, and Y. Kabashima, *Statistical mechanics of dictionary learning*, arXiv:1203.6178.
[12] Y. Iba, *The Nishimori line and Bayesian statistics*, J. Phys. A: Math. Gen. **32** (21), 3875–3888 (1999).
[13] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*, (World Sci. Pub., 1987).
[14] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, (Oxford Univ. Pr., 2001).
[15] M. Mézard and A. Montanari, *Information, Physics, and Computation*, (Oxford Univ. Press, Oxford, UK, 2009).
[16] F. Krzakala, M. Mézard, F. Sausset, Y. F. Sun, and L. Zdeborová *Statistical-Physics-Based Reconstruction in Compressed Sensing*, Phys. Rev. X **2**, pp. 021005-1–021005-18 (2012).
[17] D. J. Thouless, P. W. Anderson, and R. G. Palmer, *Solution of 'Solvable model of a spin glass*, Phil. Mag. **35** (3), pp. 593–601 (1977).
[18] K. Kabashima, *A CDMA multiuser detection algorithm on the basis of belief propagation*, J. Phys. A **36** (43), pp. 11111–11121 (2003).
[19] F. Krzakala, M. Mézard, and L. Zdeborová, *Phase Diagram and Approximate Message Passing for Blind Calibration and Dictionary Learning*. Preprint received directly from the authors via private communication.